Summary of Research

Overview

Deep Learning, and more generally AI, is addressing many challenging problems in science and engineering that were previously unsolvable (e.g., image recognition in medical science, Unstructured data analysis, AI assistants across different industries). My research focuses on developing relevant theory and algorithms for large-scale models, particularly deep neural networks. By investigating their properties as we scale them up (by increasing width, depth, data size, etc.), I aim to design computationally efficient training algorithms that are theoretically sound and practically effective. As a researcher who thrives in multidisciplinary environments, I am also interested in working with experts on domain-specific AI applications, including life sciences, education, and decision science.

1 Current Research Focus

I develop theories to understand how neural networks behave with scale, with the aim of creating novel algorithms and architectures to enhance training efficiency, stability, and performance. In addition, I am driven by the practical value of my research and am always interested in building domain-specific AI tools.

1.1 Theory and Algorithms for Large-Scale Neural Networks

A general neural network model is given by

$$Y_l(x) = \mathcal{F}_l(W_l, Y_{l-1}(x)),\ l \in \{1, 2, \dots, L\}, \tag{1}$$

where $x \in \mathbb{R}^d$ is the input, $L \geq 1$ is the network *depth* (number of layers), $(\mathcal{F}_l)_{l\in[L]}$ are mappings that define the architecture and $W_l \in \mathbb{R}^{n\times n}$ are the "weights", where n is the network *width*. To scale up the network, we can increase the width n, the depth L, the data size, the context length, etc. In this context, a key question is: How do neural network models *learn at scale*, and how can we leverage this knowledge to develop *efficient* learning algorithms? This is the main question I address in my research.

Efficient Learning at Scale. Figure 1 depicts the general approach I follow to study the learning dynamics of large-scale neural networks. This approach consists of two steps:

- **Step 1 (Theory)**: characterize the "limiting" learning dynamics of neural networks as a function of different hyperparameters in the model (activation function, learning rate, initialization etc.).
- **Step 2 (Algorithms)**: By ensuring that the limiting dynamics satisfy desirable properties (stability, feature learning, etc.), we can infer optimal hyperparameters for improved performance in large-scale settings.

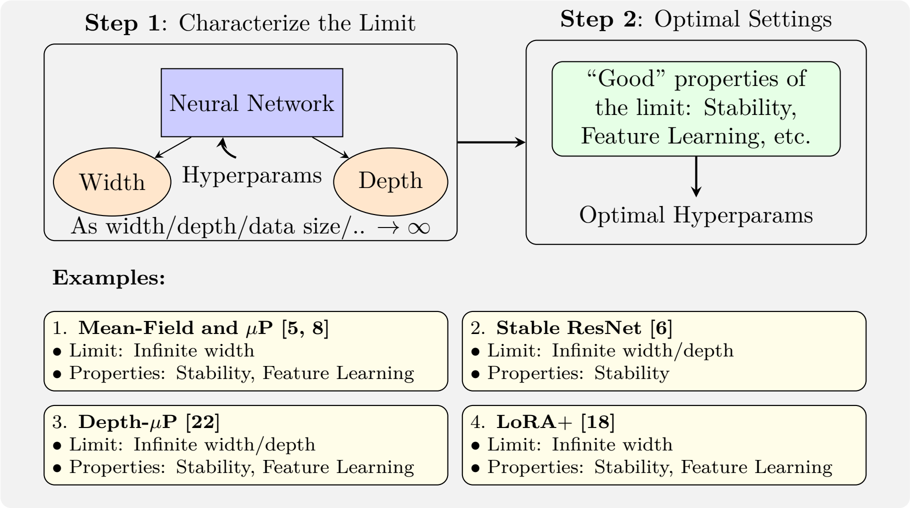
Step 1: Characterize the Limit
Neural Network
Width
Hyperparams
Depth
As width/depth/data size/.. → ∞
Step 2: Optimal Settings
"Good" properties of the limit: Stability, Feature Learning, etc.
Optimal Hyperparams
Examples:
1. **Mean-Field and μP [5, 8]**
• Limit: Infinite width
• Properties: Stability, Feature Learning
2. **Stable ResNet [6]**
• Limit: Infinite width/depth
• Properties: Stability
3. **Depth-μP [22]**
• Limit: Infinite width/depth
• Properties: Stability, Feature Learning
4. **LoRA+ [18]**
• Limit: Infinite width
• Properties: Stability, Feature Learning

Figure 1: A Framework for **Efficient Learning at Scale.**

Why study the limiting dynamics? As neural networks scale, their training dynamics tend to converge (or diverge). More importantly, the limiting dynamics are generally easier to study with existing mathematical tools. For instance, concentration results are used for infinite width analysis, dynamical systems for the infinite depth limit, etc. Below, I list some of our relevant contributions to efficient learning at scale. While most of our contributions follow the framework of efficient learning at scale by including both Step 1 (Theory) and Step 2 (Algorithms), some are limited to Step 1.

Theory: Infinite Width/Depth Neural Networks

- **Stability.** We analyzed the asymptotic behavior of neural networks as their depth increases, showing that proper parametrization leads to stable features (preactivations) in the limit. This work informs the design of networks that maintain consistent statistical properties as they scale [6, 11, 18].

- **Feature Learning.** Stability alone does not guarantee efficient learning. For instance, in the Neural Tangent Kernel regime, we obtain stability with no feature learning [2]. By examining the learning dynamics, we identified conditions under which neural networks learn meaningful features in the large-scale limit (non-trivial learning in high-dimensional settings) [22, 18].

- **Commutative Limits.** We introduced the theory of commutative scaling, showing that with certain changes in the architecture, the limits of infinite width and depth commute. This unifies the analysis of neural networks in high-dimensional regimes and provides a comprehensive understanding of their asymptotic behavior [12, 17].

Algorithms: Efficient Learning at Scale

- **Efficient Scaling (Pretraining)**: Leveraging our theoretical analysis, we developed scaling rules that stabilize gradients and maximize feature learning by appropriately adjusting the architecture and learning rates in deep networks[6, 11, 22] . This approach allows hyperparameter transfer (Depth-μP method), reduces tuning costs, and solves gradient exploding that occurs as a result of large depth [1, 6].

- **Efficient Finetuning**: Finetuning is used to boost the performance of pretrained models on specific tasks. We analyzed LoRA (Low-Rank Adaptation) and identified inefficiencies in its standard setup. We proposed LoRA+ [18], an enhanced version that corrects the learning rate configuration, resulting in consistent performance improvements. LoRA+ attracted significant community interest and is now integrated into HuggingFace's PEFT package and LLamaFactory (used by millions of practitioners). We also studied the impact of initialization on LoRA learning dynamics and proved that one initialization scheme generally outperforms alternatives [19].

- **Role of Data.** Large models are in constant need of high-quality data to improve performance. Following the efficient learning framework (Figure 1), we investigated methods for improving data quality for large models including 1) Corset Selection (data pruning), where we showed that data pruning methods fail in high compression regimes due to significant distribution shifts [10], and 2) Synthetic Data (generated from pretrained models), where we quantified the necessary proportion of original (non-synthetic) data required in a data mixture to preserve performance [20].

1.2 Real-world Challenges

I am broadly interested in Trustworthy AI, and domain-specific AI applications in life sciences and decision science.

- **Trustworthy AI.** Privacy in AI models is critical to safe deployment. In [23], we showed that smooth activation functions can increase information leakage. By analyzing signal propagation in deep neural networks, we provided a theoretical explanation for this behavior and supported our findings with empirical validations. This work contributes to understanding and mitigating privacy concerns in deep learning. Moving forward, I intend to develop methodologies in two critical areas

  1. **Reasoning**: *Create novel methodologies to analyze and understand how LLMs perform complex logical operations to improve their reasoning abilities, particularly through the Chain-of-Thought paradigm [14], which has shown impressive results with GPT4-o1.*

  2. **Hallucination**: *Investigate the causes of false or nonsensical LLM outputs, also known as hallucination [13], develop metrics to quantify hallucination, and create strategies to mitigate it.*

- **AI Applications** I am interested in exploring the potential uses of AI in real world challenges. Current areas of interest include

  1. **Life Sciences**: *Leverage AI to accelerate drug discovery [9]. Tools such as Tx-LLM [16], a finetuned model from Google's Med-PaLM 2 created to help with drug discovery, showing the potential of finetuning methods in enhancing the performance of pretrained models on such tasks.*

  2. **Education**: *I believe AI will revolutionize teaching, and I am eager to collaborate with domain experts to create personalized and adaptive learning experiences that enhance student learning outcomes [21].*

  3. **Decision Science**: *Create AI-based tools to automate decision-making processes, particularly hypothesis generation and testing. In [15], the authors showed promising results on this topic, highlighting the potential of AI in decision science.*

2 Prior Work

During my PhD, I focused on the analysis of signal propagation in deep neural networks, studying how design choices impact their performance. Our work on the “Edge of Chaos” ([4]) provided a detailed analysis of the role of the initialization and the activation function, explaining why smooth variants of ReLU (e.g. GeLU) enhance performance by enabling better information flow. We also contributed to research on the Neural Tangent Kernel (NTK) [2, 3], and network pruning, where we introduced algorithms for pruning at initialization that outperformed existing methods [7], addressing stability issues and improving the trainability of sparse networks.

References

[1] G. Yang and S. Schoenholz. “Mean field residual networks: On the edge of chaos”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 7103–7114.

[2] A. Jacot, F. Gabriel, and C. Hongler. “Neural Tangent Kernel: Convergence and Generalization in Neural Networks”. In: *Advances in Neural Information Processing Systems*. 2018.

[3] L. Chizat, E. Oyallon, and F. Bach. “On Lazy Training in Differentiable Programming”. In: *Advances in Neural Information Processing Systems*. 2019.

[4] S. Hayou, A. Doucet, and J. Rousseau. “On the Impact of the Activation function on Deep Neural Networks Training”. In: *International Conference on Machine Learning*. 2019.

[5] S. Mei, T. Misiakiewicz, and A. Montanari. “Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit”. In: *Annual Conference Computational Learning Theory*. 2019.

[6] S. Hayou, E. Clerico, B. He, G. Deligiannidis, A. Doucet, and J. Rousseau. “Stable ResNet”. In: *International Conference on Artificial Intelligence and Statistics*. 2021.

[7] S. Hayou, J. Ton, A. Doucet, and Y. Teh. “Robust Pruning at Initialization”. In: *International Conference on Learning Representations*. 2021.

[8] G. Yang and E. J. Hu. “Feature Learning in Infinite-Width Neural Networks”. In: *International Conference on Machine Learning* (2022).

[9] C. Arnold. “Inside the nascent industry of AI-designed drugs”. In: *Nature Medicine* 29.6 (June 2023), pp. 1292–1295. ISSN: 1546-170X. DOI: `10.1038/s41591-023-02361-0`. URL: `https://doi.org/10.1038/s41591-023-02361-0`.

[10] F. Ayed and S. Hayou. “Data pruning and neural scaling laws: fundamental limitations of score-based algorithms”. In: *Transactions of Machine Learning Research*. 2023.

[11] S. Hayou. “On the infinite-depth limit of finite-width neural networks”. In: *Transactions of Machine Learning Research*. 2023.

[12] S. Hayou and G. Yang. “Width and Depth Limits Commute in Residual Networks”. In: *Proceedings of the 40th International Conference on Machine Learning*. 2023.

[13] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. 2023. arXiv: `2311.05232 [cs.CL]`.

[14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. 2023. arXiv: `2201.11903 [cs.CL]`.

[15] S. Bordt, B. Lengerich, H. Nori, and R. Caruana. *Data Science with LLMs and Interpretable Models*. 2024. arXiv: `2402.14474 [cs.LG]`.

[16] J. M. Z. Chaves, E. Wang, T. Tu, E. D. Vaishnav, B. Lee, S. S. Mahdavi, C. Semturs, D. Fleet, V. Natarajan, and S. Azizi. *Tx-LLM: A Large Language Model for Therapeutics*. 2024. arXiv: `2406.06316 [cs.CL]`. URL: `https://arxiv.org/abs/2406.06316`.

[17] S. Hayou. “Commutative Width and Depth Scaling in Deep Neural Networks”. In: *Journal of Machine Learning Research*. 2024.

[18] S. Hayou, N. Ghosh, and B. Yu. “LoRA+: Efficient Low Rank Adaptation of Large Models”. In: *International Conference on Machine Learning* (2024).

[19] S. Hayou, N. Ghosh, and B. Yu. “The Impact of Initialization on LoRA Finetuning Dynamics”. In: *Advances in Neural Information Processing Systems*. 2024.

[20] M. E. A. Seddik, S.-W. Chen, S. Hayou, P. Youssef, and M. A. DEBBAH. “How bad is training on synthetic data? A statistical analysis of language model collapse”. In: *First Conference on Language Modeling*. 2024.

[21] S. Wang, T. Xu, H. Li, C. Zhang, J. Liang, J. Tang, P. S. Yu, and Q. Wen. *Large Language Models for Education: A Survey and Outlook*. 2024. arXiv: `2403.18105 [cs.CL]`. URL: `https://arxiv.org/abs/2403.18105`.

[22] G. Yang, D. Yu, C. Zhu, and S. Hayou. “Tensor Programs VI: Feature Learning in Infinite-Depth Neural Networks”. In: *International Conference on Learning Representations (ICLR)* (2024).

[23] J. Ye, A. Borovykh, S. Hayou, and R. Shokri. “Leave-one-out Distinguishability in Machine Learning”. In: *International Conference on Learning Representations*. 2024.